High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data

Benjamin Elie, Yves Laprie, Pierre-André Vuissoz, Freddy Odille

High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data

Benjamin Elie*†, Yves Laprie*, Pierre-André Vuissoz†, Freddy Odille†
*LORIA
INRIA/CNRS/université de Lorraine
†IADI, Université de Lorraine, Nancy, France
U947, INSERM, Nancy, France

***Abstract*—The paper presents a method to acquire articulatory data from a sequence of MRI images at a high framerate. The acquisition rate is enhanced by partially collecting data in the kt-space. The combination of compressed sensing technique, along with homodyne reconstruction, enables the missing data to be recovered. The good reconstruction is guaranteed by an appropriate design of the sampling pattern. It is based on a pseudo-random Cartesian scheme, where each line is partially acquired for use of the homodyne reconstruction, and where the lines are pseudo-randomly sampled: central lines are constantly acquired and the sampling density decreases as the lines are far from the center. Application on real speech data show that the framework enables dynamic sequences of vocal tract images to be recovered at a framerate higher than 30 frames per second and with a spatial resolution of 1 mm. A method to extract articulatory data from contour identification is presented. It is intended, *in fine*, to be used for the creation of a large database of articulatory data.**

Index Terms: Dynamic speech MRI, Compressed Sensing, Articulatory data

I. Introduction

Studies on speech production require the knowledge of both audio recordings and articulatory data in order to define the existing relationships between the articulatory gestures and the acoustic clues. Due to several constraints (health hazard, access to internal tissues...), the acquisition of articulatory data is challenging and raises several major issues.

Speech production involves two levels of coordination. Speech articulators, which change the acoustic properties of speech, have to move according to the sequence of phonemes to articulate. This mainly corresponds to the coarticulation phenomenon. In addition, the realization of supraglottal constrictions and cavities has to be coordinated with the glottis setup so the the aerodynamic properties of the vocal tract are compatible with the nature of the excitation source. These two coordination mechanisms have different temporal time scales. At the segmental level, i.e. that of speech sounds, the order of magnitude of the coordination between the supraglottal cavities and the vocal folds is the millisecond, while that between supraglottal articulators is the hundred of milliseconds. This means that the study of coarticulation, which is our main objective, requires a sampling rate well above 10 Hz, at least 30 Hz to get articulatory data which can be exploited. Indeed, this gives approximately one or two images at least for each speech sound.

Recently, Magnetic Resonance Imagery techniques (MRI) have been of particular interest for articulatory data acquisitions because of the numerous advantages they present [1]–[3]. Unlike X-ray images, they are harmless for the subject, and they enable a single slice to be selected, or a set of slices for a 3D representation. However, the major drawback is the required acquisition time. Acquiring the whole k-space corresponding to a single slice may take several hundred of milliseconds, and consequently, limit the temporal resolution to be no less than this value, which is not sufficiently low to correctly observe fast articulatory movements.

Historically, speech MRI have used spiral sampling schemes to enhance the acquisition rate [1]. This sampling scheme offers good image quality given the high acquisition rate. However, it may generate strong undesired artifacts, such as unrealistic tongue tip and lip elongations, that disturb the articulator contours estimation from these images. To prevent these artifacts, the proposed framework uses a Cartesian-based sampling scheme. The choice of a Cartesian-based sampling scheme is also motivated by its adaptability to be used with compressed sensing [4] and with homodyne reconstruction [5]. These mathematical frameworks allow images to be recovered from partial Fourier information. This paper presents a method to simultaneously integrate several acceleration techniques for MRI to be used in articulatory data acquisition. Sec. II details the theoretical background of compressed sensing and homodyne reconstruction applied in this paper, as well as the choice of the sampling scheme. Simulations and experimentations are presented in Sec. III, and the results in Sec. **??**. The method for exploiting articulatory data is presented in Sec. IV.

II. Theory

A. Compressed sensing

Compressed sensing (CS) is a mathematical framework enabling a subsampled signal, i.e. a signal that is sampled at a rate that does not fulfill the Shannon theory of signal sampling [4], to be recovered. The accuracy of the CS recovery lies on the assumption that the signal $\mathbf{x} \in \mathbb{C}^n$ to be recovered is K-sparse in a certain base, i.e. it exists a sparse-transform $\mathbf{\Psi}$ so that $\mathbf{\Psi x} \in \mathbb{C}^n$ contains only $K < n$ non-zero elements. In that case, in the presence of noise, only $m \geq K$ observations

of $\mathbf{x}$ are sufficient to find a solution by solving the convex problem

$$\mathbf{x} = \underset{\hat{\mathbf{x}}}{\operatorname{argmin}} ||\mathbf{\Psi}\hat{\mathbf{x}}||_1 \quad \text{s.t.} \quad ||\mathbf{\Phi}\hat{\mathbf{x}} - \mathbf{b}||_2 \leq \epsilon, \tag{1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ is the CS encoding matrix that contains only 0 and 1, so that $\mathbf{\Phi}\mathbf{x} = \mathbf{b}$, and $\mathbf{b} \in \mathbb{C}^m$ is the observed signal, namely the subsampled version of $\mathbf{x}$, and ϵ is a tolerance value to account for noise in the observation data. The ℓ_1-norm $||\mathbf{x}||_1$ of the vector $\mathbf{x}$ is defined as the sum of the modulus of the complex elements of $\mathbf{x}$. Another important condition for accurate recovery is the incoherence of the encoding matrix $\mathbf{\Phi}$ with the sparsifying transform $\mathbf{\Psi}$.

When dealing with MRI, the desired signal is usually the image intensity, namely the inverse spatial Fourier transform of $\mathbf{x}$, denoted $\mathcal{F}_{sp}^{-1}\mathbf{x}$. In our case, the sparsifying transform $\mathbf{\Psi}$ has been chosen as the temporal Fourier of the image intensity signal, namely

$$\mathbf{\Psi} = \mathcal{F}_t \mathcal{F}_{sp}^{-1}\mathbf{x},$$

where $\mathcal{F}_t$ is the temporal Fourier transfrom operator. This is justified by the fact that most of the image, namely the head, does not move, and that only a small portion makes relatively slow movements. As a result, most of the pixels exhibit a highly sparse temporal Fourier transform. In comparison with wavelet-based transforms, it has been shown to guarantee better contrasts at the vocal tract boundary walls. In practice, Eq. 1 is then

$$\rho = \underset{\hat{\rho}}{\operatorname{argmin}} ||\mathcal{F}_t\hat{\rho}||_1 \quad \text{s.t.} \quad ||\mathbf{\Phi}\mathcal{F}_{sp}\hat{\rho} - \mathbf{b}||_2 \leq \epsilon, \tag{2}$$

where ρ is the set of images to be recovered.

B. Distributed compressed sensing

In practice, MRI techniques use multichannel coils. The availability of the multichannel coils is commonly used to increase the acquisition speed by parallel imaging techniques, such as SENSE (*SENSitivity Encoding*) [6], or GRAPPA (*Generalized Autocalibrating Partially Paralleled Acquisitions*) [7]. It can also be used in CS techniques by exploiting the fact that the signals observed by the different coils are strongly correlated. Basically, their sparsity under the chosen sparse representation should be similar. Consequently, it is possible to introduce such joint sparsity as an a priori in the inverse problem by simultaneously minimizing the ℓ_1-norm of the sparse representation of the signals in each coils, and also the number of non-zero rows of the signal matrix (the matrix whose columns are the signal in each coil). This last constraint comes from the assumption that the non-zero coefficients of the sparse representation of the coil signals share the same location. This problem, called *Distributed Compressed Sensing* (DCS) [8], writes

$$\mathbf{P} = \underset{\hat{\mathbf{P}}}{\operatorname{argmin}} ||\mathcal{F}_t\hat{\mathbf{P}}||_{1,2} \quad \text{s.t.} \quad ||\mathbf{\Phi}\mathcal{F}_{sp}\hat{\mathbf{P}} - \mathbf{B}||_{2,2} \leq \epsilon, \tag{3}$$

where $\mathbf{P} \in \mathbb{C}^{n \times l}$ is the matrix whose columns contain the reconstructed image vectors of the l coils, and $\mathbf{B} \in \mathbb{C}^{m \times l}$ is the observation matrix. The mixed $\ell_{1,2}$-norm of the matrix $\mathbf{P}$ is defined as the ℓ_1-norm of the ℓ_2-norm of each row of $\mathbf{P}$, hence

$$||\mathbf{P}||_{1,2} = \sum_{i=1}^{n} ||\mathbf{P}_i||_2, \tag{4}$$

where $\mathbf{P}_i$ is the ith row of $\mathbf{P}$.

C. Homodyne reconstruction

Homodyne reconstruction techniques are based on the assumption that k-space should exhibit an hermitian symmetry because of the real-valued nature of the images to be recovered. Consequently, the knowledge of only half of the Fourier space is necessary. In practice, this property is never totally valid because of phase errors in the images. However, it is still possible to perform phase correction from a partial k-space sampling, where the portion of the sampled k-space is slightly greater than 1/2.

Homodyne reconstructions [9] use a phase correction image defined as

$$p^*(x, y) = e^{-j\angle\rho_{lr}(x,y)}, \tag{5}$$

where ρ_{lr} is a low-resolution version of the image ρ. It is the inverse spatial-Fourier transform of the observed k-space to which a rectangular weight function, centered around the very first central lines, has been applied to conserve only the low spatial frequency domain. This phase correction image is then multiplied to the image obtained by inverse spatial-Fourier transform of the zero-filled observed k-space. The reconstructed image is the the real part of the phase corrected image. Among homodyne reconstruction techniques, POCS (*Projection Onto Convex Sets*) [5] has been shown to perform well with small k-space fractions. It consists in iteratively applying the phase correction to the reconstructed data until modifications from an iteration to the next goes under the desired threshold.

In application with CS, because of large amount of missing data, the low-resolution image ρ_{lr} is computed from the temporal mean value of the acquired kt-space.

D. Choice of the subsampling pattern

Recent studies about dynamic speech MRI have focused on spiral sampling schemes [2], [3]. This sampling trajectory is very efficient for accelerating the acquisition process, but is less robust to undesired motion artifacts. Because of its simplicity and its wide use in practice, a pseudo-random Cartesian sampling scheme has been preferred for the study. Variable density sampling used with Cartesian sampling scheme, where central lines (or low frequency) are privileged has been proven to give more accurate recovery than uniformly random sampling schemes [10].

Given these considerations, the sampling trajectory is designed as follows. First, a number of encoded lines per temporal frame, called n_{lpf}, is set to an optimal value. It is a trade-off between good temporal resolution (low n_{lpf}) and high reconstruction quality (high n_{lpf}). Then, the number of fully encoded center lines, called n_{cl}, with $n_{cl} \leq n_{lpf}$, is

chosen. The n_{cl} central lines are constantly encoded at each temporal frame. Finally, the remaining $n_{lpf} - n_{cl}$ lines to be encoded at temporal frame t are randomly chosen given the following probability function

$$p(k_y, t) = \left| \frac{1}{[1 - (k_y - n_y/2)]^{r(t)}} \right|, \qquad (6)$$

where k_y is the line number of the k-space, $n_y/2$ is the center line (null frequency) of the k-space, and $r(t) \in [0, 0.5]$ is a number chosen randomly at each temporal frame according to a uniform distribution. This distribution leads to a variable sampling density that decreases as one digresses from the center lines of the k-space.

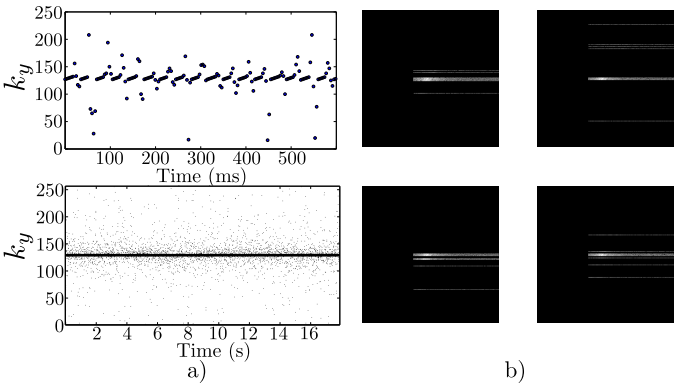

Figure 1. Sampling trajectory used for the study. Left column: sequential sampling pattern of the acquired line as a function of time: (top) detail line after line, (bottom) global pattern, frame per frame. Right : first 4 acquired k-spaces. Note that each line is not fully sampled. Missing data are recovered using the homodyne reconstruction algorithm POCS [5].

An example of sampling trajectory is given in Fig. 1. It shows the sequential pattern of the acquired line as a function of time. The first 4 subsampled k-spaces are also represented.

E. Proposed reconstruction method

To summarize, the experimental protocol to acquire articulatory data at high spatiotemporal resolution may be expressed as follows

1) choice of the slice to be recovered (usually midsagittal slice),
2) choice of acquisition parameters (n_y, n_x, n_{lpf}, and n_{cl}) according to the desired spatiotemporal resolution,
3) compute the pseudo-random Cartesian sampling scheme according to Eq. (6),
4) recover missing sampled line information with POCS algorithm [5], where the phase correction is computed from the temporal mean value of the acquired kt-space,
5) distributed compressed sensing reconstruction using Eq. (4) and SPGL1 solver [11], [12],
6) denoising processing to enhance image quality, using appropriate denoising algorithm [13],
7) coil combination to merge data into a single combined coil [14].

III. Results

A. Simulations

Simulations with numeric phantoms are performed to assess the recovery quality of the method as a function of the number of lines per frame n_{lpf}, which is related to the obtained framerate. Numeric phantoms are created from recorded images of speech MRI obtained with an independent technique [15]. The image sequences are then interpolated on a random time vector, where each time step corresponds to a single line acquisition time, to simulate acceleration and speed decrease of articulatory movements. The corresponding kt-space is then subsampled in regards with the different tested sampling trajectories. Sparse reconstruction is then applied to the simulated data. Images are 256×256 pixels of size 1.016×1.016 mm^2. Each sequence contains 128 images acquired with 16 coils. The time step for the simulation is set to the value of the used repetition time in practice, which is 3.5 ms, and n_{cl} is set to 5.

Fig. 2 shows the NRMSE (*Normalized Root Mean Square Error*) value of recovered images from the simulations as a function of n_{lpf}. To compute the NRMSE value, only pixels that are included in the region of interest defined by a rectangle bounded by the vocal tract are considered. As expected, the reconstruction quality is enhanced as the number of lines per frame increases. However, for $n_{lpf} > 10$, there is no significant increase of the reconstruction quality. Consequently, in practice, setting n_{lpf} to 10 is a good compromise between high framerate and good image quality.

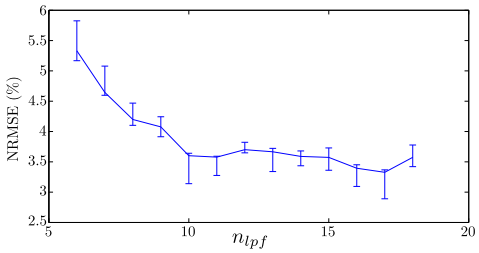

Figure 2. Median NRMSE value of recovered images from the simulations as a function of n_{lpf}. Error bars indicate upper and lower median absolute deviations, each of them computed from a set of 5 realizations.

B. MRI experiments

MRI experiments were performed on a 3T Signa HDxt MR system (General Electric Healthcare, Milwaukee, WI). Speech MRI data were obtained from 2 healthy volunteers with written informed consent and approval of local ethics committee. The data were collected with an 16-channel neurovascular coil array. The protocol consisted in a sagittal slice through the middle of the vocal track acquired with a custom modified Spoiled Fast Gradient Echo (Fast SPGR, TR 3.5ms, TE 1.1ms, line BW 83.33 kHz, flip angle 30 degrees, half echo in frequency direction, matrix 256×256, 512 temporal frames). For each temporal frame the modification consisted to acquire only a randomized fixed size sample (n_{lpf} lines) of the 256 phase lines of the k-space that gives an acquisition time

per temporal frame of $3.5\times n_{lpf}$ milliseconds and a total acquisition time for the vocal production of $512\times3.5\times n_{lpf}$ seconds.

In Fig. 3, the evolution of the lip opening as a function of time, as well as the movement of the back of the tongue during the utterance "J'ai pigé la phrase" (/ʒe.pi.ʒe.la.fʁɑz/), are plotted. Boxes in right plots indicate the production of the word "phrase" (/fʁɑz/).

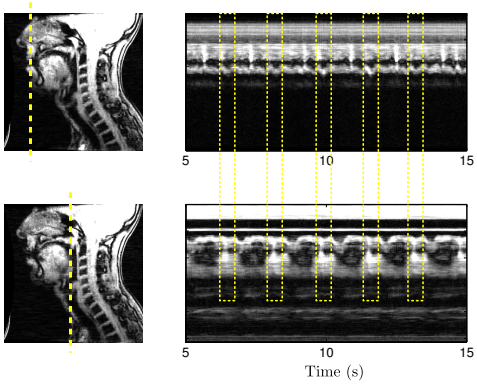

Figure 3. Strip plots of the repeated utterance "J'ai pigé la phrase" (/ʒe.pi.ʒe.la.fʁɑz/). Top: strip plot of the lip opening. Bottom: strip plot of the posterior constriction. Slice positions are indicated by a vertical dotted line on the midsagittal slice at the left. Boxes on right plots indicate the productions of the word "phrase" (/fʁɑz/). Time resolution is 35 ms. Spatial resolution is 1.016×1.016 mm^2, and slice thickness is 5 mm.

The temporal resolution and the image quality are sufficient to visualize the articulator movements. For instance, the lip opening increases to pronounce /fʁɑz/: it goes from a narrow constriction for the voiceless labiodental fricative /f/, to a relatively wide opening for the open vowel /ɑ/, and then decreases again for the alveolar voiced fricative /z/. At the same time, as shown in the bottom strip plot in Fig. 3, the back of the tongue rises to create an uvular constriction to make the voiced uvular fricative /ʁ/.

Fig. 4 shows strip plots of repeated utterance /ara/, containing the alveolar trill /r/. The recovered film has a frame-rate of 48 frames per second, and the spatial resolution is 1.016×1.016 mm^2. Moments of the production of the alveolar trill /r/ are indicated by boxes on the strip plots.

The high spatiotemporal resolution enables the oscillations of the tongue tip to be seen. It is also interesting to note the coarticulation of the tongue tip and its back: at each production of /r/, they both make simultaneous constrictions. The uvular constriction made with the back of the tongue probably enables the oral pressure upstream the alveolar constriction to raise, so that aerodynamic conditions around the tongue tip cause it to auto-oscillate.

IV. Exploiting articulatory data

Although the acoustics of the vocal tract is completely determined by the vocal tract geometry and the source conditions, outlining the contour from the glottis up to the lips as a whole does not suffice for modeling coarticulation. Indeed, articulators do not move as a whole from one position to the

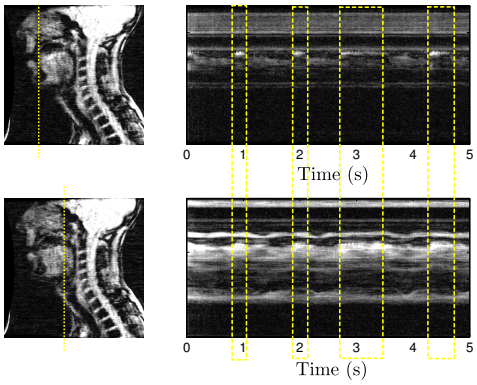

Figure 4. Strip plot of the repeated utterance /ara/. Top: strip plot of the tongue tip movement. Bottom: strip plot of the uvular constriction. Slice positions are indicated by a vertical dotted line on the midsagittal slice at the left. Boxes on right plots indicate the productions of the alveolar trill /r/. Time resolution is 20.8 ms. Spatial resolution is 1.016×1.016 mm^2, and slice thickness is 5 mm.

next. Articulators are organized in groups, each group realizing a particular gesture. Speech production involves overlapping gestures which have to be modeled separately. It is therefore important to outline each articulator independently, i.e. the mandible, the tongue, both lips, the epiglottis, the larynx and the velum.

It should be noted that bones (and particularly the mandible in the case of the vocal tract) which clearly appear on X-rays images become invisible on MR images. The mandible which gives crucial information about mouth opening has to be derived indirectly from the contour of the incisor root, thus with some imprecision due to the limited extent of this tooth on the image. Conversely, other contours and primarily the tongue, are more visible on MR images than X-ray images. We have developed specific software, called Xarticulators, for outlining articulator contours and we explored two directions for outlining contours. The first consists of delineating the tongue contour by hand. This is feasible for a small number of images in order to get good quality contours and check that acoustic simulations produce the expected acoustic signal. For larger databases it is possible to take advantage of the fact that, unlike X-ray, MR imaging produce contours which do not overlap between each other. This feature is very interesting in terms of contour tracking and we exploited it by adapting the semi-automatic tracking algorithm developed by Fontecave and Berthommier [16]. The main principle is to delineate the target contour in key images randomly selected in the cineMRI, and then to index images where the contour has to be tracked with respect to the key images by using a DCT (Discrete Cosine Transform) distance. The image used for indexing images is limited to the region (not necessarily a rectangle) where the contour to be tracked is located. The resulting contour is obtained by averaging contours of the closest images. This semi-automatic tracking requires supervision from the user who can add key images to correct errors corresponding to an image too far from the existing key images.

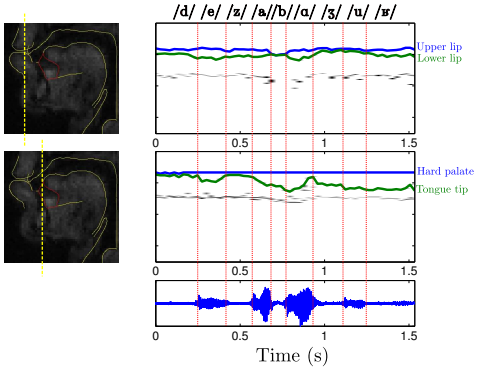

Figure 5. Strip plots of the articulator contours extracted from the utterance "Des abat-jours" (/de.za.bɑ.ʒuʁ/). Top plot: lip opening. Middle plot: alveolar region opening. Bottom plot: audio signal. Phonetic segmentation is indicated by vertical dashed lines. Left plot: midsagittal view of the vocal tract with articulator contours and position of the strip plot (vertical dotted line).

Fig. 5 shows an example of articulator contours tracked by the presented method. It corresponds to the utterance "Des abat-jours" (/de.za.bɑ.ʒuʁ/), and shows the lip opening (upper plot) and the alveolar region opening (middle plot) as a function of time. The recorded audio signal, to which MRI noise has been removed thanks to a source separation technique [17], is shown at the bottom, along with the phonetic segmentation. The temporal evolution of the articulator contours agrees with the phonetic segmentation: an alveolar constriction is created during the pronunciation of /z/, lips are closed during /b/ and suddenly opens for the following /ɑ/, and lips are very close each other for pronouncing the diphone /ʒu/.

V. Conclusion

The presented framework for acquiring 2D images of the vocal tract enables the temporal evolution of articulator contours to be extracted with a good temporal resolution. Simulations with numeric phantoms have shown that images are still accurately recovered, even with a very high decimation factor. In practice, this may degrade the image quality, and consequently, make the contour extraction more difficult, however, it could be used to analyze particular very fast movements, such as alveolar trills. In the context of natural speech, setting the framerate at around 30 frames per second is a good trade-off between image quality and acquisition rate. Movements of articulators, such as lips or tongue tip can be finely analyzed thanks to a semi-automatic tracking algorithm. An example has been presented in this paper, where tracked articulator movements agree with the phonetic segmentation of the recorded audio signal. A large articulatory database is currently being fed, following the presented framework[1]. It is intended to create accurate articulatory model from statistical methods, representing realistic anatomic data.

Acknowledgement

The authors thank FEDER and Région Lorraine for their financial support.

[1] Video files are available at:www.loria.fr/~belie/pages/csmri.html

References

[1] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the acoustical society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] Maojing Fu, Bo Zhao, Christopher Carignan, Ryan K Shosted, Jamie L Perry, David P Kuehn, Zhi-Pei Liang, and Bradley P Sutton, "High-resolution dynamic speech imaging with joint low-rank and sparsity constraints," *Magnetic Resonance in Medicine*, vol. 73, no. 5, pp. 1820–1832, 2015.

[3] Sajan Goud Lingala, Yinghua Zhu, Yoon-Chul Kim, Asterios Toutios, Shrikanth Narayanan, and Krishna S Nayak, "A fast and flexible mri system for the study of dynamic vocal tract shaping," *Magnetic resonance in medicine*, 2016.

[4] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly, "Compressed sensing mri," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 72–82, 2008.

[5] Dante C Youla, "Generalized image restoration by the method of alternating orthogonal projections," *Circuits and systems, IEEE transactions on*, vol. 25, no. 9, pp. 694–702, 1978.

[6] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, Peter Boesiger, et al., "Sense: sensitivity encoding for fast mri," *Magnetic resonance in medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[7] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase, "Generalized autocalibrating partially parallel acquisitions (grappa)," *Magnetic resonance in medicine*, vol. 47, no. 6, pp. 1202–1210, 2002.

[8] D Liang, KF King, B Liu, and L Ying, "Accelerating sense using distributed compressed sensing," in *Proc Intl Soc Mag Reson Med*, 2009, vol. 17, p. 377.

[9] Douglas C Noll, Dwight G Nishimura, and Albert Macovski, "Homodyne detection in magnetic resonance imaging," *Medical Imaging, IEEE Transactions on*, vol. 10, no. 2, pp. 154–163, 1991.

[10] F. Krahmer and R. Ward, "Stable and robust sampling strategies for compressive imaging," *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 612–622, Feb 2014.

[11] E. van den Berg and M. P. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

[12] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[13] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi, "A nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Trans. Image Process.*, vol. 22(1), pp. 119–133, 2013.

[14] Tao Zhang, John M Pauly, Shreyas S Vasanawala, and Michael Lustig, "Coil compression for accelerated imaging with cartesian sampling," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 571–582, 2013.

[15] Pierre-André Vuissoz, Freddy Odille, Yves Laprie, Emmanuel Vincent, and Jacques Felblinger, "Sound synchronization and motion compensated reconstruction for speech cine mri," in *ISMRM 2015 Annual Meeting*, 2015.

[16] J. Fontecave Jallon and F. Berthommier, "A semi-automatic method for extracting vocal-tract movements from x-ray films," *Speech Communication*, vol. 51, no. 2, pp. 97–115, 2009.

[17] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20(4), pp. 1118–1133, 2012.